Appendix

Appendix A1 Study characteristics: Chambers, Slavin, Madden, Cheung, & Gifford, 2004 (quasi-experimental design)

Characteristic	Description
Study citation	Chambers, B., Slavin, R. E., Madden, N. A., Cheung, A., & Gifford, R. (2004). *Effects of Success for All with embedded video on the beginning reading achievement of Hispanic children.* Baltimore, MD: Johns Hopkins University, Center for Research on the Education of Students Placed at Risk.
Participants	The study included 324 Hispanic students in kindergarten and the first grade.[1] Three-quarters of the students in schools involved in the study qualified for free lunch. Each of the four intervention schools were matched with similar comparison schools.
Setting	Eight schools located in high-poverty districts in the District of Columbia, New York, Arizona, California, and Illinois.
Intervention	For six months, *Reading Reels,* a series of video segments that each last up to three minutes, were shown during lessons for a maximum of six minutes. Skits and animation in the video segments demonstrate components of beginning reading and were used in conjunction with *Success for All.* The *SFA* program included a reading curriculum, tutoring, quarterly assessments, family support teams for students' parents, school personnel facilitator, and intervention teacher training. *Reading Reels* used video to teach a concept, such as defining a word and sounding it out. Students were then asked to demonstrate the concept, typically by imitating the puppets in the video sounding out a word, both individually and with partners.
Comparison	Teachers in comparison schools taught their students with the school's existing curricula.
Primary outcomes and measurement	Although four subtests of the Woodcock Reading Mastery Test were administered, only the three that require actual reading are included in this report: Word Identification, Word Attack, and Passage Comprehension. Letter Identification was excluded because it is not a measure of reading achievement. (See Appendix A2 for more detailed descriptions of outcome measures.)
Teacher training	During their first year providing *SFA* instruction, teachers participate in training for three days during the summer, and two to four additional in-service days during the school year. Throughout each year, additional in-service presentations covering classroom management, instructional pace, and cooperative learning are provided by school facilitators and other *SFA* staff. Facilitators organize information sessions in a manner that allows teachers to share problems and solutions, suggest changes, and discuss individual children. Twice a year, trainers conduct teacher observations. After the first year, training is reinforced by regular in-services, an annual *SFA* conference, and implementation checks for the facilitators and trainers. The staff development model used in the program emphasizes relative brief initial training with extensive classroom follow-up, coaching, and group discussion. Teachers also received training about showing the videos in a sample lesson.

1. Subsample data for Spanish-dominant ELL students were obtained from the study author.

Appendix A2 Outcome measures in the reading achievement domain

Outcome measure	Description
Woodcock Reading Mastery Test (WRMT): Word Identification	This subtest measures basic word reading skills. The standardized test requires the child to read aloud isolated real words that range in frequency and difficulty (as cited in Chambers, Slavin, Madden, Cheung, & Gifford, 2004).
WRMT: Word Attack	This subtest measures phonemic decoding skills by asking students to read pseudowords. For this standardized test, students are made aware that the words are not real (as cited in Chambers, Slavin, Madden, Cheung, & Gifford, 2004).
WRMT: Passage Comprehension	This subtest measures comprehension by having students fill in missing words in a short paragraph. This is a standardized test (as cited in Chambers, Slavin, Madden, Cheung, & Gifford, 2004).

Appendix A3 Summary of study findings included in the rating for the reading achievement domain[1]

Outcome measure	Study sample	Sample size (schools/ students)[3]	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Chambers, Slavin, Madden, Cheung, and Gifford, 2004 (quasi-experimental design)[8]								
WMRT: Word Identification	Kindergarten & first grade	8/324	414.29 (38.92)	404.83 (29.14)	9.46	0.26	ns	+10
WMRT: Word Attack	Kindergarten & first grade	8/324	476.59 (21.91)	467.23 (16.38)	9.36	0.45	ns	+17
WMRT: Passage Comprehension	Kindergarten & first grade	8/324	446.55 (23.03)	443.81 (20.65)	2.74	0.12	ns	+5
Domain average[9] for reading achievement (Chambers et al., 2004)						0.28	ns	+11

ns = not statistically significant

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. Subsample data for Spanish-dominant ELL students were obtained from the study author.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile ranking of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Chambers et al. (2004), corrections for clustering and multiple comparisons were needed. Therefore, the significance levels differ from those reported in the original study.
9. This row provides the study average, which in this instance is also the domain average. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size.

Appendix A4 *Success for All* rating for the reading achievement domain

The WWC rates an intervention's effects in a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.[1]

For the outcome domain of reading achievement, the WWC rated *Success for All* as potentially positive. It did not meet the criteria for positive effects because there was only one study and it did not meet WWC evidence standards for a strong design. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, negative effects) were not considered because *Success for All* was assigned the highest applicable rating.

Rating received

Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect.

  Met. *Success for All* met this criterion because the study reviewed had substantively important positive effects.

AND

- Criterion 2: No studies showing a statistically significant or substantively important *negative* effect and fewer or the same number of studies showing *indeterminate* effects than showing statistically significant or substantively important *positive* effects.

  Met. *Success for All* met this criterion because the one study reviewed did not have statistically significant or substantively important negative findings.

Other ratings considered

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *positive* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. *Success for All* did not meet this criterion because only one study was reviewed.

AND

- Criterion 2: No studies showing statistically significant or substantively important *negative* effects.

  Met. *Success for All* met this criterion because the study reviewed did not have statistically significant or substantively important negative findings.

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain-level effect. The WWC also considers the size of the domain-level effect for ratings of potentially positive or potentially negative effects. See the WWC Intervention Rating Scheme for a complete description.

Appendix A5 Extent of evidence by domain

Outcome domain	Number of studies	Sample size		Extent of evidence[1]
		Schools	Students	
Reading achievement	1	8	324	Small
Mathematics achievement	0	0	0	na
English language development	0	0	0	na

na = not applicable/not studied

1. A rating of "moderate to large" requires at least two studies and two schools across studies in one domain, and a total sample size across studies of at least 350 students or 14 classrooms. Otherwise, the rating is "small."